SEGREGATION OF QUORA INSINCERE QUESTIONS USING LSTM AND GLOVE EMBEDDING

by

ROHIT
AMAN SHAJI 17BEC1164
ADITYA HARICHANDAR A 17BEC1085

CSE4020 –MACHINE LEARNING

in

B.TECH COMPUTER AND SCIENCE ENGINEERING

Vellore Institute of Technology, Chennai
Vandalur – Kelambakkam Road
Chennai – 600127

NOVEMBER 2019

ABSTRACT

Machine learning models have been very successful when dealing with text classification problems. In this project did a measure on it tasked with classifying queries from a dataset provided by the favored web site Quora, as ‘sincere’ or ‘insincere’ . The large-scale dataset, provided by the web site Kaggle, contains over one,300,000 queries with labels to coach our models on. A separate test set that contains over 300,000 unlabeled questions is used by Kaggle to test our model. We implement models and use completely different pre-trained word embeddings to attain the most effective results.The project is based on a problem faced by quora and many other sites similar in trying to find out insincere questions among all the other genuine questions in the site. Building an algorithm using ML is what we achieved and using the dataset we were able to correctly segregate the data into insincere questions or not.We use the LSTM classification and GloVe word embeddings to do this.

Since there are various users referring to these sites for answers and trying to solve others problems, it would be in the site’s best interest to avoid anything that can reduce the sites ongoing traffic of users. And one of the big problems that causes is the ongoing increase in insincere questions typed by users. No user would like to see this in the middle of their very important research or work. That is why we decided on doing this project which will segregate the questions and help improve user experience.

We implemented this by using GloVe word embedding and using the LSTM classification after data ore-processing to do our final classification of the new dataset into insincere questions or not. Using this algorithm , the user experience is enhanced leaving the problematic questions away from us.

The algorithm was checked with the training set and predicted the accuracy rate.

The model was implemented with enough pre-processing and sufficient solutions to the problem. Future implementation on processing can be done.

I. INTRODUCTION

i. The domain of product

The domain of this project belongs in the sector of INTERNET AND SOCIETY. Since this is majorly used in segregating the insincere question the people in the society have uploaded within the internet.

Since the internet is the first thing everyone resorts to in finding new information and getting solutions to problems first hand. Quora is one the most popular sites in giving opinions and answers to questions that any party can participate. Any member can ask for solutions or opinions in the social platform and others can view and give an upvote or downvote the question. They also can answer the questions or give their own opinions about the question in hand and so on.

But due to the many users who might use this site for various propagandas and intentions that do not seem safe, the safety of user experience is drastically called into play. This is why we focused on implementing a model to solve this problem and make the user experience more safe.

Since there are various users referring to these sites for answers and trying to solve others problems, it would be in the site's best interest to avoid anything that can reduce the sites ongoing traffic of users. And one of the big problems that causes is the ongoing increase in insincere questions typed by users. No user would like to see this in the middle of their very important research or work. That is why we decided on doing this project which will segregate the questions and help improve user experience.

ii. What is the general observation about current works carried out

There are currently many works being carried out in this area. There is a whole dataset dedicated to this cause with more than a million questions segregated into binary format. There are many algorithms that has been implemented on this topic and maybe few have been made official on the site. But there is still the problem of the former where some questions passthrought the algorithms . So a best solution has not yet been found and the problem statement is still being tackled by many users alike.

iii. Organization of rest of paper

After the implementation of the model , we are using the python gui called tkinter to deliver our final product for easier access. The product will have an easy interface which takes in input from user and gives output as "VALID" or "NOT VALID".

II. BACKGROUND

Online review is the foremost and most easily accessible free info sources.These are used by both layman and bid scale organisations alike. Customers use iit to decide on buying a particular product, meanwhile organisations depend on them to improve their product. Establishments are utilizing significance of opinions to earn undue profit by hiring professionals known as spammers, giving optimistic comments on their products and pessimistic opinions on their opponent ‘s product.This activity is known as opinion spamming and should be known to give genuine results containing sentiments towards a product. So far, opinion spam detection has been advised as a discrete classification problem, generally as spam and non-spam. However, it involves incertitude as suspicious behavior of a user might be due to coincidence.

DATA PRE PROCESSING:

Before building the model, text prepossessing methods was done to clean up the data set. Empty space were removed and invisible characters were replaced with blank strings. The stopwords were deleted, and correct the most common misspelled words, i.e.‘favourite‘: ‘favorite‘. More pre processing has been done to replace strange punctuation, and clean any latex math symbols.

GLOBAL VECTORS FOR WORD REPRESENTATION(GLOVE):

GloVe is a new global log bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. It expeditiously leverages international applied math info by coaching solely on the nonzero components in an exceedingly word-word co-occurrence matrix, and produces a vector house with substantive sub-structure. It systematically outperforms word2vec on the word analogy task, given the same corpus, vocabulary, window size, and training time. It achieves better results faster, and also obtains the best results irrespective of speed.

WORD EMBEDDING:

To perform well on most natural language processing tasks , some notion of similarity and difference between words must be understood. In this project, Tfid Vectorizer and GloVe were mainly used to encode word tokens

LONG SHORT-TERM MEMORY(LSTM):

LongShort-term Memory (LSTM) is an artificial RecurrentNeural Network (RNN) field of study used in the field of machine learning and deep learning. Unlike conventional feed-forward neural networks, LSTM has response connections and other uses. It cannot solely point single knowledge points (such as images),

however conjointly entire sequences of knowledge (such as speech or video). For instance, LSTM is practical to tasks like non-segmental, connected handwriting recognition or speech identification.

A popular LSTM unit consists of a cell, associate degree input gate, associate degree output gate and a forget gate. The cell remembers values over arbitrary time intervals and therefore the 3 gates regulate the flow of data into and out of the cell.

These LSTM networks are optimal for classifying, processing and devising predictions based on time series data, since there can be holdup of unknown duration between crucial events in a time series. LSTMs were formulated to modify the increasing and disappearing gradient issues which will be encountered once training ancient RNNs. Relative inability to gap length is a plus of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

RNNs will keep track of discretional long-run dependencies within the input sequences. The problem of vanilla RNNs is machine (or practical) in nature: once training a ordinary RNN exploitation back-propagation, the gradients that area unit back-propagated will "vanish" (that is, they can tend to or become zero) or "explode" (that is, they can tend to infinity or become a very large number), because of the calculations involved in the process, which use finite-precision numbers. RNNs exploitation LSTM units partly solve the vanishing gradient downside, because LSTM units allow gradients to also flow unchanged. However, LSTM networks will still suffer from the increasing gradient downside.

TRAINING AN LSTM:

An RNN exploitation LSTM unit will be trained in an extremely supervised fashion, on a set of training sequences, using an optimisation algorithm, like gradient descent, compounded with back-propagation through time to cypher the gradients required throughout the improvement method, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with reference to corresponding weight.

A problem with exploitation gradient descent for traditional RNNs is that error gradients vanish exponentially quickly with the dimensions of the postponement between vital events. This is thanks to $\lim n \to \infty W n = \text{zero}$

However, with these LSTM units, when the error values are back-propagated from the final output layer, the error remains inside the LSTM unit's cell. This "error carousel" endlessly feeds error back to every of the LSTM unit's gates, till they learn to cut off the quality.

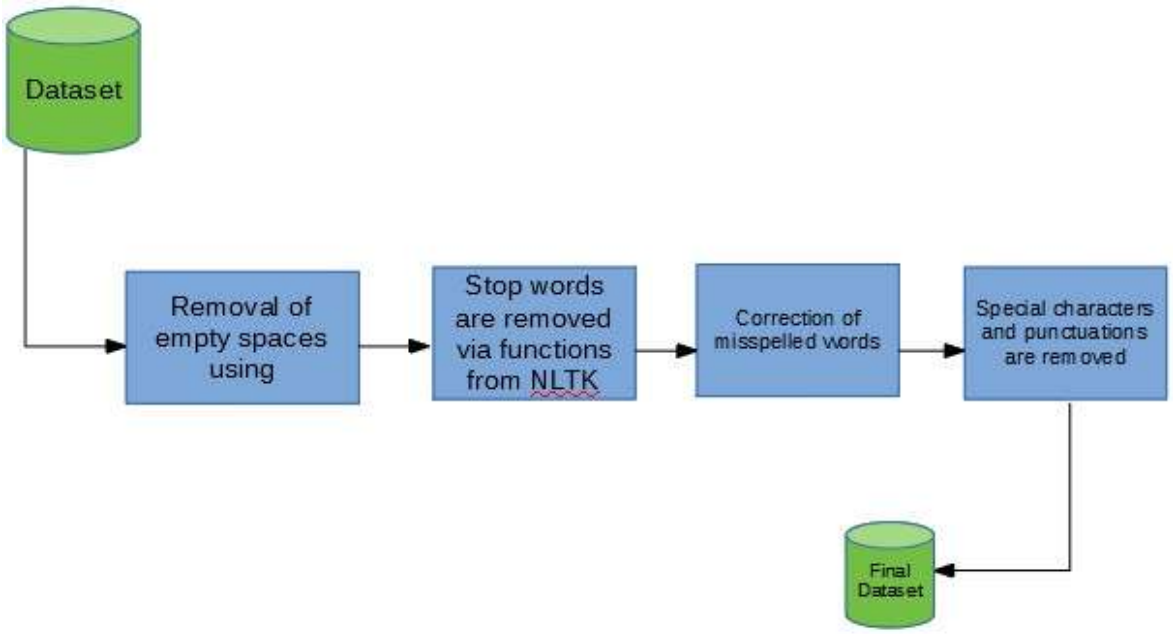
III. PROPOSED ARCHITECTURE

flowchart

DATA PRE PROCESSING:

Before building the model, text prepossessing methods was done to clean up the data set.

- Empty space were removed
- invisible characters were replaced with blank strings
- The stopwords were deleted
- Correction of the most common misspelled words, i.e.‘favourite‘: ‘favorite‘.
- More pre processing has been done to replace strange punctuation, and clean any latex math symbols.



GLOBAL VECTORS FOR WORD REPRESENTATION(GLOVE):

After pre processing and tokenizning. Glove data was loaded and the tokens were created.

GloVe is a new global log bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods

A word-word co-occurrence matrix is created along with a vector house with substantive sub-structure.

WORD EMBEDDING:

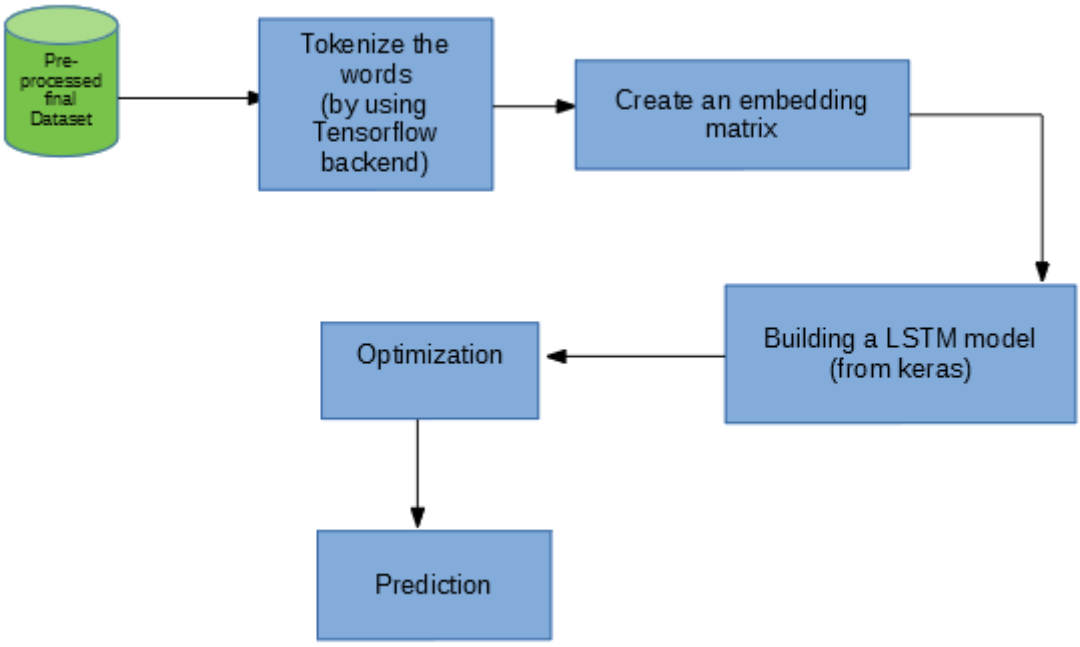
Tfid Vectorizer and GloVe were mainly used to encode word tokens

LONG SHORT-TERM MEMORY(LSTM):

LSTM networks are optimal for classifying, processing and devising predictions based on time series data

TRAINING LSTM:

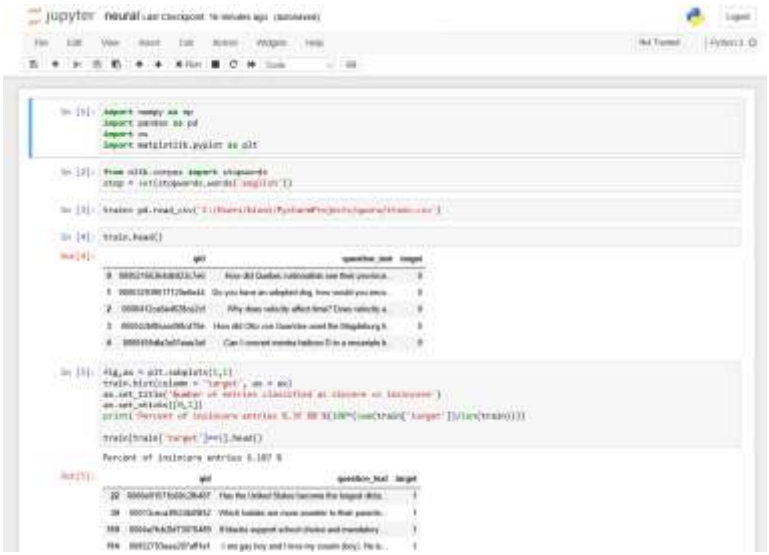
LSTM unit will be trained in an extremely supervised


Pre-processed final Dataset
Tokenize the words (by using Tensorflow backend)
Create an embedding matrix
Building a LSTM model (from keras)
Optimization
Prediction

IV. IMPLEMENTATION

- The software used is jupyter notebook with python 3 for the coding and model implementation .

The JupyterNotebook is an open-source web application which allows you to make and share documents that contain live python code, equations, visualizations and narrative text. Major uses include: data clean and transform, numerical simulation, statistical modeling of machine learning algorithms, data visualization, machine learning, and much more.



- The dataset was taken from kaggle and it contains quora sincere/insincere questions data set.

The dataset is filled with questions posted but it has no info about time of post or access other posts done by the same user. So predictions will be done just based on the text content of each question.

TEST DATA:

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0

The training dataset contains only 3 features:

- Question id : just an identity feature
- Question text : the major feature which the model will perform the fore-mentioned algorithms on
- The target: this is the output and dependent variable.

1:INSINCERE QUESTION

0:SINCERE QUESTION

TEST DATA:

	qid	question_text
0	0000163e3ea7c7a74cd7	Why do so many women become so rude and arroga...
1	00002bd4fb5d505b9161	When should I apply for RV college of engineer...
2	00007756b4a147d2b0b3	What is it really like to be a nurse practitio...
3	000086e4b7e1c7146103	Who are entrepreneurs?
4	0000c4c3fbe8785a3090	Is education really making good people nowadays?

- The GUI is done using the tkinter, which is one of the most popularly used GUIs. It has standard python interfacing and it also outputs the fastest and easiest way to create and simulate GUI applications

V. RESULT AND DISCUSSION

```
In [4]: train.head()
```

Out[4]:

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0

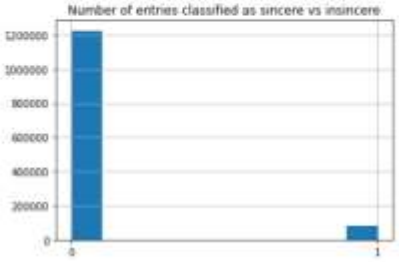
```
In [5]: fig,ax = plt.subplots(1,1)
train.hist(column = 'target', ax = ax)
ax.set_title('Number of entries classified as sincere vs insincere')
ax.set_xticks([0,1])
print('Percent of insincere entries %.3f %%'%(100*(sum(train['target'])/len(train))))

train[train['target']==1].head()
```

```
Percent of insincere entries 6.187 %
```

Out[5]:

	qid	question_text	target
22	0000e91571b60c2fb487	Has the United States become the largest dicta...	1
30	00013ceca3f624b09f42	Which babies are more sweeter to their parents...	1
110	0004a7fcb2bf73075488	If blacks support school choice and mandatory ...	1
114	00052793eaa287aff1e1	I am gay boy and I love my cousin (boy). He is...	1
115	000537213b01fd77b58a	Which races have the smallest penis?	1



On seeing the train dataset we were able to find out that about 6 percent of the data currently take from Quora was insincere

```
In [25]: result = df1.join(df2)
result
```

Out[25]:

	qid	pred
0	0000163e3ea7c7a74cd7	1
1	00002b4ffb5d505b9161	0
2	00037756b4a147d2b0b3	1
3	000086e4b7e1c7146103	1
4	0000c4c3fbe8785a3090	1
5	000101884c19f3515c1a	1
6	00010f6253778f1f44a47	1
7	00012afbd27452239059	1
8	00014894849d00ba98a9	1
9	000156468431f09b3cae	1
10	00015c487b65d0f79cd8	1
11	0001880504d9d091c8c8	1
12	00019b780e31adab8acd	1
13	0001d5bbc5f7b7a5ae58	1
14	0001f64b6aaf396c4cc4	1
15	00022773433360e1aaea	1
16	00022821df8a68a75586	1
17	0002564a23df8d58e6ff	1
18	00031573ca46574f518c	1
19	0003165642ba9bd3539	1
20	000354e56f1f937c9be6	1
21	000383e4ebd22f6ade04	1
22	0003a7d231d5036979d4	1

On seeing the predicted values we see tthat most of the test data is predicted and insincere questions and in fact the only sincere question was the second one in almost 20 questions

VI. CONCLUSION AND FUTURE WORKS

Our future work will have three aspects.

• Pre-trained Model:Trysomepre-trainedmodel,likethe pre-trainpartofBERTforwordembedding,itmaybenefit thefeatureextraction. • Attention: Add attention to our neural network and proposeanAttention-basedLongShort-TermMemoryNetwork for aspect-level sentiment classification.The attention mechanism can concentrate on different parts of a sentence when different aspects are taken as input.

• CNN:Use convolutional neural networks can be used as a recurrent structure to capture contextual information as far as possible, which may introduce considerably less noise compared to traditional window based neural networks.

VII. REFERENCES

- Using Word2Vec to process big text data
  Publisher: IEEE
  https://ieeexplore.ieee.org/abstract/document/7364114
- Sentimental Analysis and opinion mining
  Publisher:Bing Liu
  University of Illinois at Chicago
  https://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016
- The Diffusion of News Applying Sentiment Analysis and Impact on Human Behavior Through Social Media
  https://link.springer.com/chapter/10.1007%2F978-3-030-20454-9_25
- The Word2Vec Algorithm
  https://www.datasciencecentral.com/profiles/blogs/the-word2vec-algorithm
- Spam analysis of big reviews dataset using Fuzzy Ranking Evaluation Algorithm and Hadoop
  https://link.springer.com/article/10.1007%2Fs13042-017-0768-3
- An Introduction to Recurrent Neural Networks
  https://medium.com/explore-artificial-intelligence/an-introduction-to-recurrent-neural-networks-72c97bf0912
- Deep Recurrent Neural Networks for Hyperspectral Image Classification
  https://ieeexplore.ieee.org/document/7914752
- Air Pollution Forecasting Using RNN with LSTM
  https://ieeexplore.ieee.org/document/8512020